"A method for determining the biological likelihood of candidate compositions or structures"

Cross-Reference to Related Applications

The present application claims priority from Provisional Patent Application No 2003905362 filed on 1 October 2003, the contents of which is incorporated herein by reference.

Field of the Invention

This invention relates to a method of determining the biological likelihood of candidate compositions or structures, particularly glycans and their derivatives also known as oligosaccharides. As used hereinafter the term glycan will include both glycans and glycan derivatives, unless otherwise indicated.

Background of the Invention

Glycans (sugar structures/oligosaccharides) are usually composed of varying numbers of less than a dozen biologically-occurring monosaccharides. When considered purely in terms of their masses there are usually only about 3-6 different mass-unique monosaccharides in a typical glycan structure. The most frequently encountered unique-mass monosaccharides are Hex (mass 162 Da; includes all hexose monosaccharides), HexNAc (mass 203 Da; includes all acetamidohexose monosaccharides), dHex (mass 146 Da; includes all deoxyhexose monosaccharides), Pent (mass 132 Da; includes all pentose monosaccharides), and NeuAc (mass 291 Da; N-acetylneuraminic (sialic) acid). There are several other biologically extant, though less-frequently encountered component monosaccharides, such as KDN, HexA, NeuGc. Other non-monosaccharide adducts such as sulfate (S; mass 79.97 Da), phosphate (P; mass 79.97 Da), methyl (14 Da), and acetyl (42 Da) are also occasionally observed on biologically-occurring oligosaccharides.

It is often the case during the characterisation of biological molecules that a precise mass may be ascertained for each biological molecule but its composition and identity are unknown. Given a reasonably accurate mass, such as would normally be obtained by mass spectrometry, the monosaccharide composition of an unknown glycan can be theorised by determining, by computation, the set of monosaccharide compositions that are within a reasonable mass deviation (or tolerance) of the observed mass. This approach forms the basis of glycomod (http://us.expasy.org/tools/glycomod/) a publicly available research tool. The shortcomings of this tool, and of this purely theoretical approach by extension, is that a large number of compositions are returned for any mass of larger than moderate size, and that the majority (90-99%) of these have little in common with known biologically extant compositions.

The aim of the present invention is to attempt to alleviate some of the above described problems and to reduce the large number of irrelevant compositions returned by existing tools.

Any discussion of documents, acts, materials, devices, articles or the like which has been included in the present specification is solely for the purpose of providing a context for the present invention. It is not to be taken as an admission that any or all of these matters form part of the prior art base or were common general knowledge in the field relevant to the present invention as it existed before the priority date of each claim of this application.

Summary of the Invention

In a first broad aspect, the present invention incorporates statistical measures of biological relevance for the candidate compositions returned.

Typically, biological relevance, expressed as a numerical score, or biological index, is determined by statistical comparison to an established reference set of known and fully characterised compositions, in the case of glycans a reference set such as the Glycosuite (http://www.glycosuite.com) database. The biological index of any given composition may then be used as a basis for discarding biologically "unlikely" compositions, as well as for ranking (sorting) of returned compositions by biological likeliness.

Empirically, for glycans this allows between 90-99.9% of candidate compositions returned by any given search to be discarded, whilst preserving and ranking the remaining, biologically likely compositions.

In one aspect the present invention provides a method of determining the likelihood of a candidate composition comprising:

selecting a reference group of known characterised compositions;

establishing statistical characteristics relating to components of or other features of the known characterised composition;

comparing the statistical characteristics of the known characterised compositions with corresponding components or features in the candidate compositions to establish a likelihood of those compositions occurring.

More particularly, for glycans, in one aspect the present invention provides a method of characterising glycans comprising the steps of:

providing a search mass of a glycan whose composition is to be determined;

generating a list of candidate glycans, typically including theoretical glycans, made up of components, including monosaccharides, whose total mass is within a predetermined tolerance of the search mass;

selecting a reference group of known characterised glycan compositions of approximately similar mass to the search mass;

establishing the mean and standard deviation of each component appearing in the reference group of the known characterised glycan compositions;

for each candidate glycan composition calculating a partial score for each component in that glycan candidate based on the difference between the observed number of the component in the glycan candidate and the mean for that component in the reference group, divided by the standard deviation;

combining the partial scores to provide an indication of the likelihood of that theoretical glycan candidate occurring.

The candidate glycans will include structures which exist as well as theoretically possible glycan structures which are not known to exist.

The predetermined tolerance of the search mass is within +/- 400Da, preferably +/- 200Da.

The partial scores may be combined in any suitable manner. One way, for example, is by multiplying the partial scores together.

By the use of actual biological information, the present invention is able to discern biologically likely compositions from the vast majority of compositions of similar mass, but whose compositions are differ greatly from known, biologically extant compositions. For example, for glycans where the publicly available web tool glycomod returns over 100 theoretical compositions for the mass 1300 Da +/- 0.5 Da, a tool embodying the present invention returns 2 biologically likely compositions and 109 biologically unlikely compositions (which would normally be discarded).

Although the main application of the present invention is to the delineation of biologically likely and unlikely sugar compositions for the purposes of sugar structure/composition elucidation, the generic methodology of using known biological data as a means to refine, interpret, and/or rank theoretical or empirical data may be used for many other applications.

Brief Description of the Drawings

Specific examples of the present invention will now be described by way of example only with reference to the accompanying drawings in which:

Figure 1 illustrates a computer running software embodying aspects of the present invention;

Figure 2 shows the data entry page of a software based search tool embodying the invention;

Figure 3 shows a results page of the search tool of Figure 2;

Figure 4 shows the data entry page of the software based search tool for a mass that is known not to exist in the Glycosuite database;

Figure 5 shows the results page for the search presented in Figure 4;

Detailed Description of a Preferred Embodiment

The present invention is implemented on a computer means running software carrying out the algorithms and process of the method. The computer is illustrated in Figure 1 and comprises a processor or CPU 100, a visual display screen 102, keyboard 104, mouse 106 and printer 108. The computer is connected to a database 110 (known as GlycoSuite) by the internet a LAN or the like

The first input to a search using a method to determine a glycan composition (the "search glycan") embodying the present invention, is a search mass (which is typically in Daltons). The search mass is typically an empirically determined mass of the "search glycan" which is to be characterised determined by mass spectrometry or other means i.e. the mass of the search glycan whose composition is to be determined.

A search mass tolerance (in Daltons) is also input. Typically this will be a relatively small value depending on the expected accuracy of the empirically determined search mass and typically may be of the order of ± 0.1Da. Also input is a "biological index" cut-off. The biological index is a measure of a theoretical glycan composition's likelihood and its derivation is explained in more detail below. The cut off is the value of that index above which candidate compositions are discarded as being too unlikely to occur in the real world. Also input is a "maximum composition" which indicates the maximum allowable number of each monosaccharide in each theoretical glycan composition. By way of example, if it were known for a fact that the glycan to be characterised contained no sialic acid, the theoretical glycan compositions generated as potential matches for the search mass would also exclude sialic acid. This reduces the amount of computation required and improves speed and accuracy. In the system implementing the method, defaults would typically be provided for those inputs, except of course for the search mass.

Other optional parameters may also be exposed to the user to further modify the performance of the search. The output of the composition search is a list of candidate glycan compositions, the majority of which will be theoretical compositions, i.e. possible structures but ones which are not known to be extant, whose mass is within the search mass tolerance of the search mass, and whose biological index is less than the biological index cut-off. In theory one of those candidates matches the composition of the search glycan. The list of candidate glycan compositions may include naturally occurring glycan compositions.

The composition search is performed as follows:

Reference statistics for the given search mass are determined from the (Glycosuite) database. This process is described in more detail below.

Monosaccharides are recursively recombined in varying numbers such that every possible combination of allowed monosaccharides is created. Compositions whose mass does not fall within the search mass tolerance are discarded, as are compositions for which the number of any monosaccharide exceeds the maximum number of that monosaccharide specified by the "maximum composition". The result is a list of theoretical candidate glycan compositions.

The biological index of candidate compositions is determined as described below. Compositions whose biological index does not satisfy the biological index cut-off are discarded.

The remaining compositions are presented to the user in order of biological index. Typically the list will be short and may only include one or two candidates. This compares with the hundreds of candidates typically produced by Glycomod, each of which has to be individually reviewed and assessed.

Calculation of Biological Index

Inputs to the process are a composition, and a reference data set of known sugar compositions/structures. The reference set may be from any suitable database or data source such as Glycosuite. The output of the process is a numerical biological index.

The determination of biological index for a given search glycan composition proceeds as follows:

The mass of the composition is the search mass or may be determined by the sum of the residue masses of each monosaccharide/component in the composition.

By reference to the reference set of known biological compositions, the mean and standard deviation of every monosaccharide/component in the database within an arbitrary mass range (eg: +/- 200 Da) of the mass of the composition is determined. Obtaining statistics from a range of masses around the given composition's mass is necessary in order to obtain a sufficiently large sample size (preferably at least 100 known compositions). In the case of the Glycosuite database of known sugar structures, a mass tolerance of 200 Da was empirically determined to be sufficient to provide in excess of 100 known compositions for search masses up to around 3500.

By way of example if the search mass were 1000 Da there may be 100 known glycans in the database whose mass is between 800 and 1200 Da. The mean and standard deviation of each of every monosaccharide/component appearing in those known glycans in the database is then determined. If we take HexNAc as an example we may find that, on average, the 100 known glycans contain 3.3 HexNAc monosaccharides with a standard deviation of 2.3. This process is repeated to calculate the mean and standard deviation for each monosaccharide component Hex, dHex, pent et al, and each adduct in the known glycans, if adducts are being accounted for.

For each candidate glycan composition "Partial scores" are then determined from the means and standard deviations calculated above. These are calculated for each monosaccharide in the given composition as the absolute value of the difference between the mean number of that monosaccharide in the reference set and the observed number of that monosaccharide in the theoretical candidate composition, divided by the standard deviation of that monosaccharide in compositions from the reference set. ie:

$$partialscore_{monosac} = \frac{|mean_{monosac} - observed_{monosac}|}{stdev_{monosac}}$$

where $mean_{monosac}$ is the mean 'number of the given monosaccharide in the reference data set (Glycosuite); $mean_{monosac}$ is the number of the given monosaccharide in the theoretical candidate composition; and $stddev_{monosac}$ is the standard deviation of the given monosaccharide in the reference data set.

By way of example if the theoretical glycan composition includes two HexNAc, three Hex and 1 NeuAc, the partial score for each of those three monosaccharides is calculated for that theoretical candidate glycan composition. Partial scores need not be calculated for monosaccharides which do not appear in the candidate theoretical glycan composition.

In the event that the $mean_{monosac}$ equals the $mean_{monosac}$ for a particular glycan, the system is arranged to give the partial score a minimum value of 0.01.

Thus, the partial score of a monosaccharide is in fact the number of standard deviations the number of away from the mean that that monosaccharide is in the candidate composition. In a normal distribution, approximately 68% of all data points lie within 1 standard deviation of the mean, ~93% within 2 standard deviations, over 99% within 3. Assuming that the distributions of monosaccharide number for the mass range used to obtain the initial means and standard deviations for the given search mass are sufficiently close to normal, then partial scores of 3 or less for any monosaccharide indicate that the number of those monosaccharides are within 99% of all compositions of similar mass in Glycosuite.

Partial scores are then combined in some manner to derive a single numeric score; this being the biological index. The actual mathematical derivation of the biological index may be arrived at using multiple means; different formulae exhibit subtlely different qualities in their sensitivity to large partial scores and other criteria. For this reason, biological index for the purposes of the present invention may be considered merely as a numerical value that is representative of, and derived from, the magnitudes of the differences between a given composition and a population of known compositions of a similar mass. Presently, a biological index is calculated from partial scores as the product of all the partial scores from the candidate composition; ie:

$$BI = \prod_{monosac_0}^{monosac_n} partialscore_{monosac}$$

The Biological Index is adept at excluding very poor matches but at the same time if a candidate glycan composition has a very large (i.e. poor) partial score for one monosaccharide but low partial scores for the other monosaccharide components, the candidate may have an acceptably low Biological Index hence the system does not discard candidates which have only one poor partial score.

The process of calculating the partial scores is carried out for each candidate glycan composition as discussed above. Compositions whose biological index does not satisfy the biological index cut-off are discarded. The remaining compositions are presented to the user in order of biological index. Typically the list will be short and may only include one or two remaining candidates. This compares with the hundreds of candidates typically produced by Glycomod, each of which has to be individually reviewed and assessed.

The key element of the present invention is the use of biological data as a means to score the quality of theoretical data, in this case, sugar compositions. The actual manner in which a biological score/index is calculated is largely arbitrary; different formulae for calculating a biological index exhibit different characteristics with respect to their tolerance to large compositional differences from the determined mean, and in their propensity to extrapolate the compositions present in the reference database.

Example 1

Figure 2 shows an entry page 200 of a software based search tool embodying aspects of the present invention. Three masses 202 are entered. In the example they are experimentally determined from bovine alpha-2-HS glycoprotein (fetuin). The entry page also allows the entry of tolerance 204, a biological index cut-off value 206 and a maximum composition for various glycan components.

Figure 3 shows the results page 300 for the search presented in Figure 2, listing all theoretical oligosaccharides compositions that are within the specified tolerance of each of the given search masses. Compositions that are present in the Glycosuite database are presented as hyperlinks to their relevant Glycosuite record and are underlined. Candidate compositions are presented in descending order of biological index. Compositions with a biological index of 2 or less are regarded as being very biologically likely. Those with a biological index of 2 to 10 are less likely but are considered to be sufficiently close to Glycosuite compositions of comparable mass. In the example, the search tool has identified the correct oligosaccharide composition and ranked it as the most biologically likely in each case. Biologically unlikely compositions (i.e. those having a biological index of greater than 10) are discarded and are not shown.

Example 2

Figure 4 shows the data entry page of the software based search tool for a mass 400 that is known not to exist in the Glycosuite database.

Figure 5 shows the results page 500 for the search presented in Figure 4. Three candidate oligosaccharide compositions are shown at 502, 504, 506, for the search mass 2943.4Da. The correct oligosaccharide composition is the third suggested composition 506. All three suggested compositions have very low biological indexes indicating that all three compositions are deemed biologically likely, despite the fact that none of the three compositions is actually present in the Glycosuite database. This demonstrates that the searching tool can effectively extrapolate the Glycosuite database. Note that without using the biological index scoring and cut-off, there are 23 candidiate compositions that would satisfy the given (non-lenient) search criteria. With lenient search parameters (i.e. unrestricted limits on monosaccharides), the number of theoretical compositions is over 100.

Example 3

Reduced oligosaccharides released from a human respiratory mucin were detected by LC-MS. The compositions of the oligosaccharides were determined manually from MS/MS data, and the measured masses were submitted to searching by the software based search tool embodying the present invention(as known as "GlycoComp"). The masses were also submitted to GlycoMod (a different glycan searching tool) for comparison. No limits were placed on the possible numbers of monosaccharides, except for methyl and acetyl groups, which were set to zero. Four of the compositions are present in the GlycoSuite database, while the largest mass is not present.

m	Correct Composition					**GlycoComp Rank**	Possible Compositions	
	HexNAc	Hex	dHex	NeuAc	Sulf		GlycoComp (B.I. < 100)	GlycoMod
531.20	1	1	1			**1**	2	2
734.28	2	1	1			**1**	3	3
976.29	2	2	1		1	**2**	14	19
1041.37	2	2		1		**1**	9	15
2866.04	5	5	3	2		**3**	122	1736

Using a biological index cut-off value of 100 resulted in a reduction in the total number of likely compositions for each mass. This is particularly notable with large masses. The "GlycoComp" rank, sorted by biological index, also makes selection of the correct structure much easier, especially when compared with GlycoMod results, which are sorted only by mass.

Variations to the described method are possible. For example, the partial score calculation as outlined above may be varied for improved accuracy. This improved monosaccharide partial score PartialScore_m first calculates the number of standard deviations from the mean for each monosaccharide:

$$\text{StDevScore}_m = \text{Abs}(\text{count}_m - \text{mean}_m)/\text{stdev}_m$$

This number of standard deviations from the mean is then converted to a probability, based on the normal distribution, for each monosaccharide:

$$\text{PartialScore}_m = \frac{e^{-\frac{1}{2}(StDevScore_m)^2}}{\sqrt{2\pi} \times stdev_m}$$

This improved partial score then represents the probability, based on similarity to the group of known oligosaccharides, that a particular number of monosaccharides are in fact present in a candidate oligosaccharide. The Biological Index is then the product of the partial scores for all monosaccharides. This method of calculating the Biological Index provides a combined total probability for the candidate composition. The biological index is then transformed to a larger number for easier interpretation.

$$\text{BiologicalIndex} = \frac{1}{\ln\left(\prod_{m \in monosaccharides} PartialScore_m\right)}$$

Partial scores are calculated based on the average characteristics of oligosaccharides present in GlycoSuiteDB within +/- 200 Da of the measured oligosaccharide mass. This molecular weight filter is based on the assumption that oligosaccharides with similar masses are more likely to have similar monosaccharide compositions. However, other filters based on biological source, disease state, or specific protein, could also be applied based on related assumptions – for example, that oligosaccharides from blood proteins, from homo sapiens, with cancer may have similar monosaccharide compositions. These filters could be user-defined in various combinations.

Filters could also include more processed measures, such as the ratios of monosaccharides in an oligosaccharide (eg fucose:hexose, HexNAc:all monosaccharides). Rather than averaging values after a particular filter, equations of best fit could also be used to calculate partial scores (eg the ratio of fucose:hexose as a function of molecular weight). These could also be used in various combinations (eg the ratio of fucose:hexose is a function of molecular weight, with a filter for only oligosaccharides from mammals). Variance associated with these lines of best fit (eg R^2 values) could be used to calculate partial scores.

Since oligosaccharides from the same biological source are likely to be structurally and compositionally related, a nested biological index (NBI) could be used to determine how similar the best ranked compositions, based on biological index, are to each other. This might be particularly useful to suggest cases in which a composition has a high biological index, due to similarity to oligosaccharides in the GlycoSuite database, but is quite different to the compositions returned for other search masses from the same sample. Since experimental oligosaccharides from one sample are likely to have a wide range of masses, a mass filter is unlikely to be useful in this case. However, other measures, such as ratios of monosaccharides, may be less dependent on molecular weight. The best ranked compositions for each mass could constitute a 'database', to which each individual composition could be compared. This comparison could be made by a NBI, or other analysis, such as principal component analysis (PCA). If there were any matched compositions (by biological index) which were significantly different to the other matches, this could indicate that they were not the correct composition, based on the biological characteristics of the sample. Compositions ranked lower by biological index could then be included in place of the putatively incorrect matches, and NBIs again calculated. This could proceed until there were minimal matches that were different to other returned matches. (eg until the standard deviations of each monosaccharide for all matched compositions were minimised).

A further extension to the calculation of biological index would be to incorporate established rules into the calculation of probabilities. Calculating the percentage of structures within a window that conform to a particular rule and converting this to a percentage could achieve this.

To automatically account for glycopeptide mass additions to an oligosaccharide, the software based search tool could either digest a known glycopeptide with a protease, or add the amino acid residue masses to the software based search tool's alphabet in order to predict possible amino acid compositions in addition to oligosaccharide compositions. The number of possible amino acid compositions may be limited by requiring the presence of certain amino acid motifs that sugars can attach to.

Although the present invention as described above is concerned with the use of known sugar structures/compositions as a means to discern/elucidate monosaccharide composition given only a mass, the concept could be extended to other compositions and to the use of other structural characteristics, for example linkage and branching, as reference data for determining and/or ascertaining the quality of complete sugar structures for other investigative techniques, such as glycan fragment mass fingerprinting (see the applicant's co-pending provisional patent application No 2003902907, the entire contents of which are incorporated herein by reference).

It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the invention as shown in the specific embodiments without departing from the spirit or scope of the invention as broadly described. The present embodiments are, therefore, to be considered in all respects as illustrative and not restrictive.